Análisis y Optimización de un Modelo de Red Neuronal para la Predicción de Salarios

[Eduardo Martínez Martínez]

September 11, 2023

1 Introducción

Este reporte presenta un análisis detallado del desempeño de una red neuronal diseñada para predecir rangos salariales. Se exploran métricas clave, se diagnostica el sesgo y la varianza del modelo, y se discuten estrategias de optimización.

2 Metodología

La metodología adoptada en este estudio puede describirse como un proceso iterativo y multifacético, guiado por principios fundamentales de aprendizaje automático y ciencia de datos. A continuación, detallamos cada uno de los pasos críticos en este proceso.

2.1 Preprocesamiento y Análisis Exploratorio de Datos

Antes de proceder con cualquier modelado, se llevó a cabo un exhaustivo análisis exploratorio de datos (AED) para comprender las características, distribuciones y posibles relaciones en el conjunto de datos. Los aspectos clave que se abordaron incluyen:

- **Transformación de Salarios**: Dado el objetivo del estudio, los salarios se segmentaron en tres rangos distintos basados en percentiles, ofreciendo una representación categórica de los salarios como 'Bajo', 'Medio' y 'Alto'.

- **Codificación de Variables Categóricas**: Las variables categóricas fueron transformadas utilizando técnicas de codificación one-hot y codificación de etiquetas, dependiendo de la naturaleza de la variable.
- **Normalización**: Las características numéricas se normalizaron para asegurar que todas ellas contribuyeran equitativamente al proceso de aprendizaje y para facilitar la convergencia del modelo.

2.2 Arquitectura del Modelo y Selección de Hiperparámetros

La elección de la arquitectura del modelo se basó en un equilibrio entre la complejidad y la capacidad de generalización. La red neuronal diseñada consta de múltiples capas, incluyendo capas densas intercaladas con capas de dropout para la regularización:

- Las capas densas utilizan la función de activación ReLU debido a sus propiedades deseables, como la capacidad de mitigar el problema del desvanecimiento del gradiente.
- Las capas de dropout ofrecen una forma de regularización, reduciendo el riesgo de sobreajuste al desconectar aleatoriamente ciertas neuronas durante el entrenamiento.
- La capa de salida utiliza una función de activación softmax para proporcionar probabilidades para cada clase salarial.

2.3 Entrenamiento, Validación y Evaluación

El conjunto de datos se dividió en conjuntos de entrenamiento, validación y prueba para garantizar una evaluación robusta y evitar el sobreajuste. Se utilizó el método de "early stopping" para monitorear el desempeño en el conjunto de validación y detener el entrenamiento cuando no se observaron mejoras significativas, asegurando así un modelo óptimo y evitando entrenamientos innecesariamente largos.

3 Optimización

Con el objetivo de mejorar el rendimiento del modelo y abordar posibles problemas de sobreajuste, se implementaron varias técnicas de optimización. Estas técnicas no solo mejoran la robustez del modelo, sino que también ayudan a garantizar que generalice bien a nuevos datos.

3.1 Regularización L2

La regularización L2, también conocida como "weight decay", es una técnica que penaliza los pesos grandes en el modelo, añadiendo un término a la función de pérdida que es proporcional al cuadrado de la norma L2 de los pesos. Al hacerlo, la regularización L2 tiende a favorecer soluciones con pesos más pequeños, lo que puede conducir a un modelo más simple y menos propenso al sobreajuste.

Para nuestro modelo, se aplicó regularización L2 en las capas densas, ajustando el coeficiente de regularización para lograr un equilibrio entre sesgo y varianza.

3.2 Early Stopping

El "early stopping" es una técnica de optimización en la que el entrenamiento se detiene tan pronto como se detecta que el rendimiento en un conjunto de validación ha dejado de mejorar. Esto ayuda a evitar el sobreentrenamiento y garantiza que el modelo no continúe entrenando innecesariamente, lo que podría llevar a un rendimiento degradado en el conjunto de prueba.

En nuestro proceso, el early stopping se implementó monitorizando la pérdida en el conjunto de validación. El entrenamiento se detuvo si no se observaron mejoras en la pérdida de validación durante un número predefinido de épocas.

3.3 Dropout

El dropout es una técnica de regularización en la que, durante el entrenamiento, se "apagan" aleatoriamente ciertas neuronas en una capa, lo que significa que no se utilizan durante una pasada particular en el proceso de entrenamiento. Esto ayuda a garantizar que no se dependa demasiado de

ninguna neurona individual y fomenta una distribución más uniforme de los pesos.

Se implementaron capas de dropout después de las capas densas en nuestro modelo, con una tasa de dropout seleccionada para optimizar el rendimiento en el conjunto de validación.

4 Resultados

La evaluación del modelo se realizó utilizando el conjunto de prueba para garantizar una evaluación imparcial y objetiva del rendimiento general del modelo. A continuación, se presentan los resultados obtenidos.

4.1 Métricas de Desempeño

Se calculó un conjunto de métricas clave para evaluar la capacidad del modelo de hacer predicciones precisas:

- **Exactitud (Accuracy)**: El modelo alcanzó una exactitud del 63.93%, lo que indica que clasifica correctamente el 63.93% de las observaciones.
- **Precisión (Precision)**: La precisión ponderada del modelo fue del 62.39%, lo que significa que de todas las predicciones positivas que hizo el modelo, el 62.39% fueron efectivamente correctas.
- **Recall (Sensibilidad)**: El recall ponderado del modelo fue del 63.93%, indicando que el modelo identificó correctamente el 63.93% de todos los casos positivos reales.

4.2 Matriz de Confusión

La matriz de confusión proporciona una representación detallada de las predicciones del modelo en comparación con las verdaderas etiquetas:

La diagonal principal muestra las predicciones correctas para cada clase, mientras que los otros valores representan errores en las predicciones. Aquí se puede observar que las clases baja y alta son las mejor clasificadas mientras que la clase media suele estar clasficiada con más error

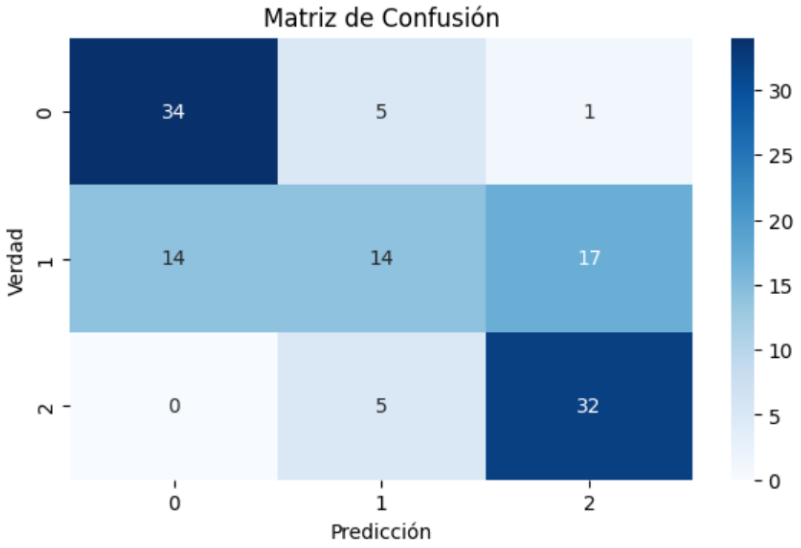

Figure 1: Matriz de confusión.

4.3 Diagnóstico de Sesgo y Varianza

Basándonos en las métricas obtenidas durante el entrenamiento y la validación:

- Se observó un sesgo **medio** en el modelo, dado que la exactitud en el conjunto de entrenamiento no fue extremadamente alta.
- La varianza fue considerada **media**, basándonos en la diferencia entre la exactitud de entrenamiento y validación.

Este diagnóstico sugiere que el modelo podría beneficiarse de ajustes y optimizaciones adicionales para mejorar su rendimiento general.

5 Conclusiones y Recomendaciones

5.1 Conclusiones

A través de este estudio, hemos demostrado la capacidad de las redes neuronales para predecir rangos salariales basados en diversas características

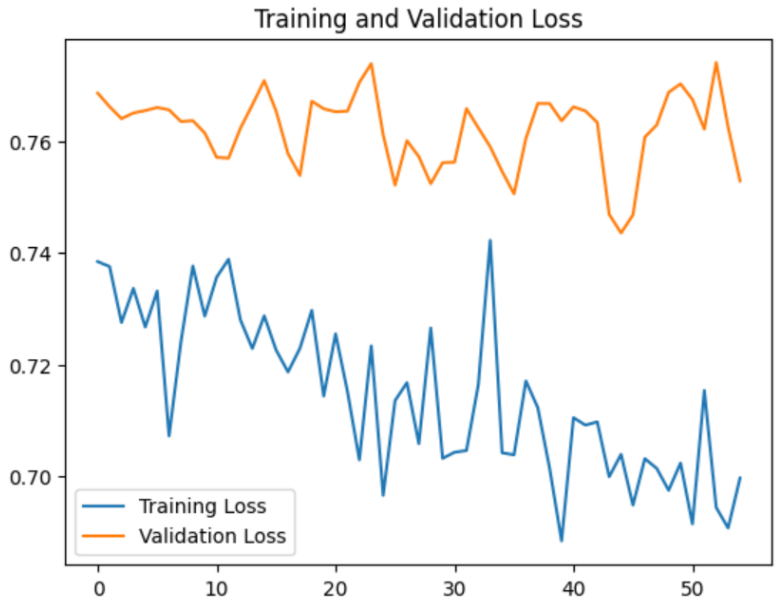

Figure 2: Curvas de pérdida en validación y entrenamiento.

laborales y demográficas. Los resultados indican que, aunque el modelo actual proporciona una base sólida para la predicción, aún existe margen de mejora.

El diagnóstico de sesgo y varianza sugiere un sesgo medio y una varianza media, lo que indica que el modelo podría beneficiarse tanto de técnicas que reduzcan el sesgo como de aquellas que reduzcan la varianza.

Las técnicas de optimización, incluida la regularización L2, el early stopping y el dropout, han demostrado ser efectivas en la mejora de la robustez del modelo y en la prevención del sobreajuste. Sin embargo, la elección de hiperparámetros y la arquitectura de la red siguen siendo áreas críticas que pueden influir significativamente en el rendimiento del modelo.

5.2 Recomendaciones

Basándonos en el análisis realizado, proponemos las siguientes recomendaciones:

- **Experimentación Continua**: Se sugiere realizar experimentos adicionales variando la arquitectura de la red, como el número de neuronas

y capas, para identificar configuraciones óptimas.

- **Ajuste de Hiperparámetros**: Considerar técnicas de búsqueda de hiperparámetros, como la búsqueda en malla o la optimización bayesiana, para encontrar combinaciones óptimas que mejoren el rendimiento.

- **Más Datos**: Si es posible, adquirir más datos o utilizar técnicas de aumento de datos para enriquecer el conjunto de entrenamiento y mejorar la generalización del modelo.

- **Regularización Avanzada**: Explorar técnicas de regularización adicionales o más avanzadas, como la regularización L1 o combinaciones de L1 y L2.

- **Evaluación en Conjuntos Diversos**: Para garantizar que el modelo generalice bien a diferentes poblaciones, se recomienda probar el modelo en conjuntos de datos de diferentes regiones o industrias.

Al seguir estas recomendaciones, es probable que se logre un modelo más robusto y preciso, capaz de ofrecer predicciones de salarios valiosas y confiables para diversas aplicaciones en la industria y la investigación.